第 43 卷 第 2 期
2023 年 4 月
桂林电子科技大学学报
Journal of Guilin University of Electronic Technology
Vol. 43, No. 2
Apr. 2023

基于稀疏优化的异常分布检测方法

陈启超，李　宽

(东莞理工学院 网络空间安全学院，广东 东莞　523808)

摘　要：现代神经网络可能会对来自训练分布之外的输入产生高置信度的预测结果，对机器学习模型构成潜在威胁。检测异常分布的输入是在现实世界中安全部署模型的核心问题。基于能量模型的检测方法，直接利用模型提取的特征向量计算样本的能量分数，而依赖并不重要的特征可能会影响检测的性能。为了解决该问题，提出了一种基于稀疏优化的损失函数。对已经预训练完成的分类模型进行微调，在学习过程中保持模型分类能力的同时，增加正常样本特征的稀疏程度，使得正常样本的能量分数降低，正常样本与异常样本之间的分数差异变大，从而提高检测效果。该方法并未引入异常的辅助数据集，避免了样本之间相关性的影响。在数据集 CIFAR-10 和 CIFAR-100 上的实验结果表明，该方法将检测 6 个异常数据集的平均 FPR 95 分别降低了 15.02%和 15.41%。

关键词：神经网络；异常分布检测；能量分数；微调；稀疏优化

中图分类号：TP391.4　　**文献标志码**：A　　**文章编号**：1673-808X(2022)05-0131-08

Sparsity-based regularization for out-of-distribution detection

CHEN Qichao, LI Kuan

(School of Cyberspace Security, Dongguan University of Technology, Dongguan 523808, China)

Abstract: Modern neural networks may produce high confidence prediction results for inputs from outside the training distribution, posing a potential threat to machine learning models. Detecting inputs from out-of-distributions is a central issue in the safe deployment of models in the real world. Detection methods based on energy models directly use the feature vectors extracted by the model to calculate the energy score of a sample, and reliance on features that are not significant may affect the performance of the detection. To alleviate this problem, a loss function based on sparse regularization is proposed to fine-tune a classification model that has been pre-trained to increase the sparsity of in-distribution sample features while maintaining the classification power of the model during the learning process. This results in a lower energy score for in-distribution samples and a larger difference in scores between in-distribution and out-of-distribution samples, thus improving detection performance. Furthermore, the method does not introduce an external auxiliary dataset, avoiding the effect of correlation between samples. Experimental results on datasets CIFAR-10 and CIFAR-100 show that the method reduced the average FPR 95 of detecting the six abnormal datasets by 15.02% and 15.41% respectively.

Key words: neural network; out-of-distribution detection; energy score; fine-tuning; sparsity regularization

近年来，在机器学习领域基于深度神经网络(DNN)的算法在分类任务上取得了显著的成果[1]。在设计算法阶段，处于静态和封闭的理想环境中，使用特定分布的数据集合训练 DNN 分类模型，针对特定分布输入数据并做预测。在真实应用场景中，模型容易遇到异常分布的输入，而它也会尝试去做预测，对这类输入的预测结果可能会出乎意料。实际上，越来越多的工作证明[2-3]，DNN 模型存在对异常分布数据过拟合的问题，即对于来自未知分布的输入数据，可能会产生一个过度自信的错误结果。

在现实世界中部署机器学习模型时，确保系统的可靠性和安全性至关重要。一个可靠的机器学习系统不仅应该对在其熟悉分布的已知输入上产生准确的预测，还应该能检测到其不熟悉分布的未知输入并

收稿日期：2022-03-31
基金项目：国家自然科学基金(61876038)
通信作者：李宽(1984—)，男，副教授，博士，人工智能、机器学习中的安全问题。E-mail：likuan@dgut.edu.cn
引文格式：陈启超，李宽. 基于稀疏优化的异常分布检测方法[J]. 桂林电子科技大学学报，2023，43(2)：131-138.

拒绝，随后再将它们移交给人类用户进行安全处理。例如，在无人驾驶汽车中[4]，希望驾驶系统能够在检测到以前从未见过的异常场景或者物体，在无法做出安全决定时，可及时发出警报并迅速将驾驶控制权移交给人类。为了在真实世界的应用场景中安全地部署 DNN 模型，异常分布检测问题被提出并迅速引起广泛关注[5]，其旨在确定模型的输入是否来自训练数据的分布，以防止模型产生不可靠的预测结果。异常分布检测要求模型不仅需要准确处理看得见的类(通常称为正常样本)，还需要能够有效处理看不见的类(异常样本)。从实践的角度来看，随着基于 DNN 的算法在工业界广泛应用，使得模型具备这种能力成为一种必然要求。Liu 等[6]提出一种能量评分函数并将其用于检测异常分布，为进一步提高检测效果，还提出了一种基于能量的微调模型的损失函数。在学习过程中，该函数需要用异常数据集来辅助模型进行训练，并将相对较低的能量值分配给正常训练数据，将较高的能量值分配给辅助训练的异常数据，从而达到正常样本能量值低、异常样本能量值高的目的。然而，该方法在训练阶段引入了依赖于数据集的超参数，且由于大规模的辅助异常数据集，训练成本(包括时间和存储消耗)特别高。

鉴于此，提出了一种基于稀疏优化的方法，用于改进基于能量的方式。该方法不引入额外数据集和超参数，易于实现和使用，可显著降低训练成本，并避免数据集相关性带来的问题，从而提高基于能量分数的检测性能。

1 基于预训练模型的方法

基于预训练模型的方法是指给定一个已预训练好的 DNN 分类模型，利用其输出空间中的信息来计算样本的异常分数，从而检测出异常数据，优点是易于使用，无需修改训练程序和目标。Hendrycks 等[5]提出基于预训练模型的检测方法 MSP，使用模型对输入数据产生的分类进行预测，即用 Softmax 概率向量中概率的最大值作为异常分数，以区分待测试样本是否异常。

Liang 等[7]对 MSP 方法进行了改进，提出了 ODIN 方法。该方法对 Softmax 层的输出按照比例进行缩放，并在待测数据的输入空间中添加一定扰动，以增加正常样本与异常样本之间预测概率的间隔，提高了检测性能。Hsu 等[8]提出了广义的 ODIN，通过分解模型输出的预测概率的方式扩展了 ODIN，避免了需要使用到异常数据的问题。Mahalanobis[9]先用预训练模型提取正常样本的特征，再用类条件高斯分布对该特征进行建模，计算待测试样本与高斯分布之间的马氏距离，并将其作为异常分数。EBD[6]从能量模型的视角推导出了一种能量评分函数，认定具有较低能量的测试样本为正常样本，具有较高能量值的测试样本为异常样本，该方法提升了能量分数的检测效果。

1.1 基于异常值暴露的方法

基于异常值暴露的方法是指借助一些多样化的，将与正常样本、测试样本具有不同分布的样本作为辅助模型训练的“异常值”，在训练期间将其主动暴露给模型，鼓励模型去学习正常样本与异常样本之间的差异，进而启发模型去概括和检测看不见的异常。最早的异常值暴露是 OE[10]，其在分类任务的基础上，通过引入一个大规模的异常样本数据集[11]，提出了一个额外的优化目标。优化过程中，模型需要在正确分类正常样本的同时，对给定的辅助异常样本产生均匀分布的 Softmax 预测概率。此后，Papadoroucos 等[12]注意到通过额外的置信度校准可为 OE 带来改进。Yu 等[13]提出一种双分支的分类网络，在训练过程中通过最大化 2 个分类器决策边界之间的距离来增加正常样本与异常样本的可区分性。当无可用的异常样本时，可用算法合成异常样本来增大不同分布之间的距离。Lee 等[14]利用 GAN 生成异常训练样本，在低置信度区域生成边界样本，并强制模型将其预测为均匀分布，但这类方法训练成本很高。

为了有效利用异常样本，可通过使用异常值挖掘[15]或对抗重采样[16]的方法来获得紧凑且具有代表性的数据集合；考虑更实际的场景，由于给定的异常样本中可能包含正常样本，可用伪标记[17]或者过滤正常样本的方法[18]来减少引入的数据相关性对模型的干扰。一般来说，基于异常值暴露的方法可达到更好的检测效果，但异常值暴露对于异常数据集的可用性强加了强有力的假设，这在实践中可能是不可行的。在现实环境中，由于客观因素不可避免地会引入正常样本，过滤掉的成本很高。鉴于此，提出一种基于稀疏优化的方法，无需用到异常数据集辅助模型训练。

2 基于稀疏优化的检测框架

2.1 问题描述

异常分布检测可以形式化为一个二元分类问题。假设 P_X 为任意样本空间 X 上定义的数据分布，从 P_X 中独立同分布采样一批样本，作为训练模型的数

据集合 D_{in}，其中样本所对应的标签集合 $Y=\{1,2,\cdots,C\}$。在异常分布检测领域的上下文中，通常把与 D_{in} 同分布的样本称为"正常样本"，否则为"异常样本"。异常样本是指来自不相关分布的样本，其标签集合与 Y 无交集。

使用训练数据集 D_{in} 训练一个 DNN 分类模型 f。异常分布检测的目标是通过设计一个估计函数 $G(x;f)$ 来评估输入的数据是否来自分布 P_X。检测器 $G(x;f)$ 定义如下：

$$G=\begin{cases}0, & S(x;f)\leqslant\delta;\\ 1, & S(x;f)>\delta。\end{cases} \tag{1}$$

其中：$S(x;f)$ 为计算待测试样本异常程度的评分函数；δ 为检测器的阈值。

阈值 δ 根据实际的需求来选择，用以调节模型的容错能力，在实验中，通常选择能够将 95%的正常样本正确识别情况下所对应的值。若置信度分数 $S(x;f)$ 大于 δ，则检测器 $G(x;f)$ 认为待测试样本为正常样本，标记为"1"；反之，$S(x;f)$ 小于 δ，则认为待测试样本为异常样本，标记为"0"。

2.2　能量评分函数

能量模型[19]（energy based model，简称 EBM）最早由 Lecun 等提出。假设存在函数映射 $F(x):\mathbf{R}^D\rightarrow\mathbf{R}$，将样本空间中的输入 x 映射成为一个被称之为"能量"的标量值，一组能量值 $F(x,y)$ 的集合可通过吉布斯分布表达为概率密度的形式：

$$p(y\mid x)=\frac{\mathrm{e}^{-F(x,y)}}{\int_y \mathrm{e}^{-F(x,y)}}=\frac{\mathrm{e}^{-F(x,y)}}{\mathrm{e}^{-F(x)}}。 \tag{2}$$

其中：$\int_y \mathrm{e}^{-F(x,y)}$ 是归一化常数，通常也被称为配分函数；$F(x)$ 为自由能，

$$F(x)=-\log\int_y -\mathrm{e}^{F(x,y)}。 \tag{3}$$

在 DNN 架构中，任给一个分类模型 $f(x):\mathbf{R}^D\rightarrow\mathbf{R}^K$，将样本空间中的输入 x 映射成为 K 维的特征向量，通常被称之为 logit，可利用 Softmax 函数计算出样本被预测为某个类的概率：

$$p(y\mid x)=\frac{\mathrm{e}^{f_y(x)}}{\sum_i^K \mathrm{e}^{f_i(x)}}。 \tag{4}$$

通过式(2)、(4)建立能量模型与 DNN 之间的联系。从能量模型的视角，DNN 中的能量函数可定义为 $F(x,y)=-f_y(x)$，当 $f(x)$ 的参数固定时，可用 Softmax 函数的分母表示能量函数：

$$F(x;f)=-\log\sum_{i=1}^K \mathrm{e}^{-f_i(x)}。 \tag{5}$$

通过 DNN 输入的 logit 可计算样本的能量值，基于能量的检测方法使用式(5)作为待测样本的评分函数 G，将具有较低能量的测试样本认为正常样本，高能量值样本为异常样本。MSP 评分函数在计算样本分数时只关心最大值的元素，忽略了其他位置元素中蕴含的信息；而 G 评分函数考虑到所有位置的元素，在一定程度上可避免过拟合问题。

2.3　稀疏优化的学习框架

给定一批有标签的正常样本训练数据集 $D_{in}=\{X,Y\}$，训练一个卷积神经网络(CNN)分类模型，损失函数采用分类中最常见的交叉熵(CE)损失：

$$L_{ce}=-\frac{1}{|D_{in}|}\sum_{(x_i,y_i)}\log(p_{y_i}(y\mid x_i))。 \tag{6}$$

模型训练完成后，得到一个具有分类正常样本能力的预训练模型 $f_\theta(x)$。此模型也具备一定的检测异常样本的能力，可通过能量分数直接计算待测试样本的分数，从而检测异常样本。但是，该模型对异常样本的检测能力往往不足，因为训练过程中专注的是分类的任务。

为进一步提升模型检测异常样本的能力，Liu 等[6]引入了一个大规模的辅助数据集，提出了一个额外的学习目标。训练过程中迫使模型为正常样本分配更低的能量值，为异常样本分配更高的能量值，从而增加正常样本与异常样本之间的能量差。具体地，能量边界的损失函数

$$L_{energy}=E_{(x_{in},y)\sim D_{in}}(\max(0,F(x_{in})-m_{in})^2+E_{(x_{out})\sim D_{out}}(\max(0,m_{out}-F(x_{out}))^2), \tag{7}$$

其中：D_{out} 为引入的大规模辅助异常数据集；$E_{(x_{in},y)\sim D_{in}}$ 和 $E_{(x_{out})\sim D_{out}}$ 分别为在 D_{in} 和 D_{out} 两个数据集上的经验风险；$F(x_{in})$ 和 $F(x_{out})$ 分别表示正常样本和异常样本的能量值；m_{in} 和 m_{out} 分别表示与 2 个数据集相关的超参数。

分类正常样本是模型的基本能力。在微调模型过程中，既要保证不能损坏模型的分类能力，也要达到增加正常样本与异常样本能量差值的目标。因此，需要将二者结合起来优化。总的目标函数

$$L_{tune}=L_{ce}+\lambda\cdot L_{energy}, \tag{8}$$

其中：L_{ce} 为训练分类模型时采用的交叉熵损失；$\lambda>0$ 为平衡能量的超参数。

容易看出，上述的能量边界损失函数的设计相当复杂，不仅引入了大规模的数据集，还引入了 2 个与训练数据集绑定的超参数，在实践的过程中，可能难以复现其效果，也很难迁移到其他场景。因此，提出

了一种基于稀疏优化的微调模型。如果进一步增加正常样本特征向量的稀疏性,而对异常样本的特征向量无限制,则正常样本的能量值可能会进一步降低,基于能量的检测算法也可以更好地检测出异常样本。图 1 为基于稀疏优化的微调模型框架。

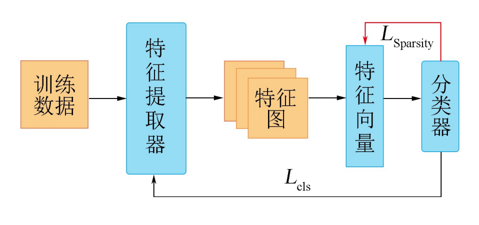


图 1　基于稀疏优化的微调模型框架

预训练完成的 CNN 分类模型 $f_\theta(x)$ 包含一个特征提取器,可提取输入的特征,输出特征图像,随后经过全连接层,输出特征向量。用所提方法对特征向量进行稀疏化处理,提出了一个稀疏正则项

$$L_{\text{sparsity}}=\frac{1}{|D_{\text{in}}|}\sum_{x_i\in D_{\text{in}}}(\|f(x_i)\|_1),\qquad(9)$$

其中,$\|f(x_i)\|_1$ 表示特征向量的 L_1 范数。

同样,微调模型不能损害模型的分类能力。在保持分类效果的同时,再进一步增加正常训练样本特征向量的稀疏度。因此,所提方法总的优化目标函数

$$L_{\text{tune}}=L_{\text{ce}}+\lambda L_{\text{sparsity}},\qquad(10)$$

其中:L_{ce} 为训练分类模型时采用的交叉熵损失;$\lambda>0$ 为平衡稀疏程度的参数。

对比式 7、式 9 可以发现,所提方法可大大降低目标函数的复杂度。与引入 2 个与数据集相关超参数的能量有界损失函数不同,稀疏优化损失是一种几乎无超参数的方案。训练过程指导模型为正常样本生成更稀疏的特征向量,无需约束异常样本的特征。所提方法未引入大规模的辅助异常数据集,只需正常样本训练数据集来微调模型即可,更加容易实现,也易于迁移。

3　实验

3.1　实验数据集

选择分类任务中常用的 CIFAR-10 和 CIFAR-100 数据集[20]作为正常样本训练数据,用于训练深度神经网络的图像分类模型,其皆由 50 000 张训练图像和 10 000 张测试图像组成,是自然场景下的 RGB 彩色(3×32×32)图像。CIFAR-10 包含了 10 个类别的数据,均是一些生活中常见的物体,如猫、狗、飞机等。CIFAR-100 包含了 100 个类别,具有更丰富的语义信息和更加具体的语义标签,样本包含细粒度和粗粒度 2 种标签。例如,在 CIFAR-10 中样本的标签是“车”,而在 CIFAR-100 中样本不仅有“车”的粗粒度标签,还会具体到是什么车,如“皮卡车”的细粒度标签。

为保证实验结果的准确,对评估实验的 OOD 数据集进行了严格把控,确保待测试的 OOD 数据集、训练模型的 ID 数据集及暴露给模型的异常数据之间,在样本的语义信息上能够尽量做到基本不交叉,即模型未提前见过待测数据。由于数据集的规模问题,很难保证一定完全无交叉,比如异常样本的局部信息可能包含了正常样本情况,人为很难精确控制。

选择 6 个不同的测试数据集来评估所提方法,这些数据集涵盖了现实世界中绝大部分常见的场景,包括自然图片风景数据集 Places365[21]、街道场景编号数据集 SVHN[22]、生活中常见纹理的数据集 Texture[23]、街道场景和物体数据集 iSUN[24],大规模场景理解数据集中根据目标信息进行裁剪得到的数据集 LSUN-Crop、缩放得到的数据集 LSUN-Resize[25]。在评估实验的过程中,所有图像都被下采样到 32×32 大小,与正常样本尺寸保持一致。

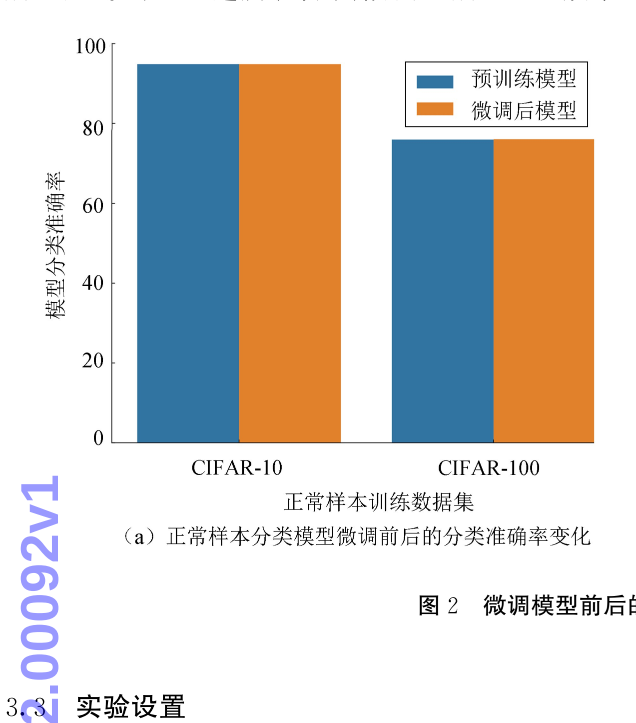
3.2　评价指标

评价指标遵循异常检测研究中的标准实验设置,采取 3 个度量指标来衡量方法的有效性,分别是:FPR 95,AUROC,AUPR。记 TP,TN,FP,FN 分别表示真阳性、真阴性、假阳性和假阴性,则正阳率 TPR=TP/(TP+FN),表示预测为正常样本且真实也为正常样本占所有真实情况中正常样本总数的比率;而假阴率 FPR=FP/(FP+TN),表示预测为正常样本但真实情况为异常样本占所有真实情况中异常样本总数的比率;查全率 Recall=TP/(TP+FN),其定义与 TPR 相同;查准率 Precision=TP/(TP+FP),表示被正确识别到的正常样本占所有真正的正常样本总数的比率。

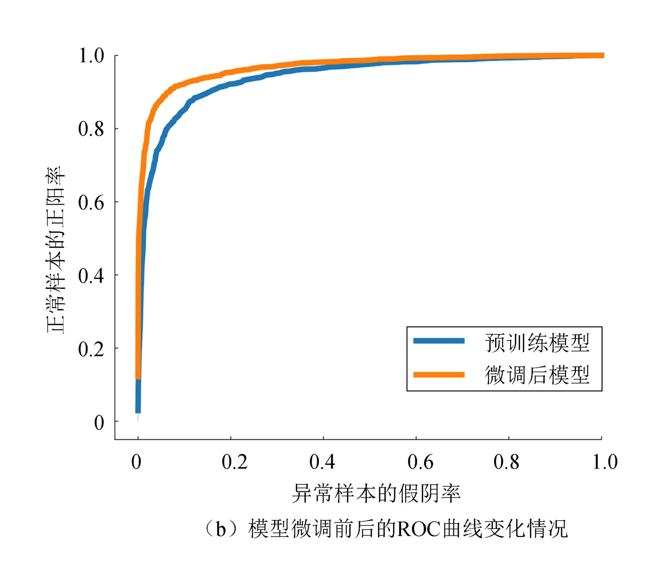
FPR 95 表示当 95%的正常样本都能够被模型正确识别的情况下所对应的 FPR;ROC 曲线为以 FPR 为横坐标、TPR 为纵坐标所对应点绘制出来的曲线,AUROC 表示 ROC 曲线与坐标轴围成区域的面积;同样,PR 曲线为以查全率为横坐标、查准率为纵坐标所对应点绘制出来的曲线。AUPR 表示 PR 曲线与坐标轴围成区域的面积。FPR 95 的值越小越好,而 AUROC 和 AUPR 的值越大越好。

图 2(a)为模型在两个正常样本数据集上,微调模型前后对正常样本分类准确率的变化情况,可发现所提方法对正常样本分类的能力几乎没有影响。而

图 2(b)为一组 ROC 曲线变化情况,其中蓝色曲线表示模型微调之前,橙色曲线表示所提方法微调模型之后,可以发现,经过所提方法微调之后,FPR 减小,TPR 增大,ROC 曲线朝着左上方移动,AUROC 的值增加。



(a)正常样本分类模型微调前后的分类准确率变化



(b)模型微调前后的ROC曲线变化情况

图 2 微调模型前后的 ROC 和分类准确率变化

3.3 实验设置

采用具有残差结构的 WideResNet[26] 作为 CNN 骨干网络来训练正常样本的分类模型,在数据预处理阶段,使用随机翻转和随机剪裁的数据增强方法,以增加数据集的多样性。此外,根据 CIFAR-10 和 CIFAR-100 数据集每个通道(彩色图像有 RGB 三色通道)的标准均值和方差进行归一化处理,以消除奇异数据产生的不良影响。训练过程中,CNN 网络的参数保持一致,每次循环加载的 BatchSize 是 128;优化算法采用随机梯度下降;优化过程采用 Momentum 动态调整梯度和权重衰减策略,Momentum 的值设为 0.9;L_2 范数的权重衰减系数设为 0.000 5,稀疏平衡常数 λ 设置为 0.01。

值得注意的是,预训练阶段与微调阶段采取的学习率和训练的 epoch 不同,预训练时学习率取值相对较大,因为训练是从一个随机的参数空间中搜索最优解,较大的学习率有助于学习在一个大的范围中迅速更新参数;而微调阶阶段采用的是一个已经具备一定知识的模型,此时选择较小的学习率,在局部的空间中搜索解,只需要少次的训练即可。具体地,在预训练模型阶段,epoch 设置为 100,学习率设置为 0.1;在微调模型阶段,其他配置保持不变,epoch 设置为 10,学习率设置为 0.001。

3.4 实验结果分析

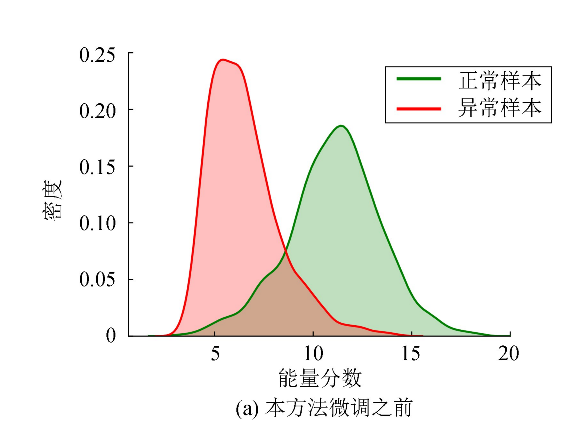
图 3(a)、(b)分别为本方法模型微调前后,所有给定待测试样本能量分数的分布,其中绿色曲线表示正常样本的分布,红色曲线表示异常样本的分布,2 条曲线交叉的部分表示正常样本与异常样本混淆的区域,在该范围内正常样本和异常样本无法区分。交叉的区域越小,说明模型的检测能力越强。微调后模型交叉的区域显著减小,说明所提方法显著提高了模型的检测能力。结合图 2(a)为分类准确率变化情况,说明所提方法在提高模型检测异常样本能力的同时,并不会损害模型对正常样本的分类能力。图 3(a)、(b)中横坐标和纵坐标所对应的比例不一样,这是由于微调后得到的是一个新模型,每个待测试样本重新输入新模型,此时提取的每个样本特征都已经发生了变化,对应的分数自然会发生改变;而在检测异常样本,并不关心具体数值,只需找到一个能够把二者区分开的阈值即可。

3.4.1 对比不同异常数据集

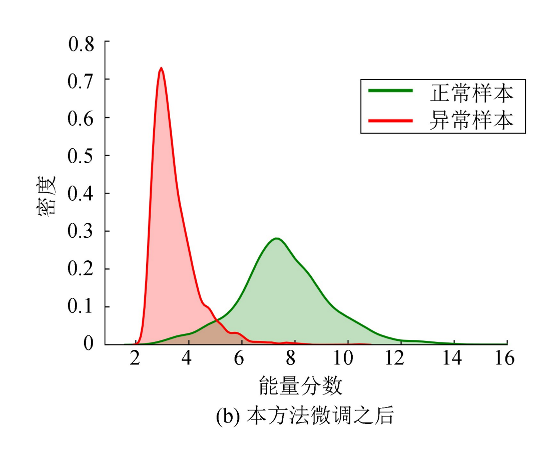
实验以 WideResNet 训练 CIFAR-10 和 CIFAR-10 的分类模型为基模型,以测试基模型的检测效果作为基准;随后,用所提方法进一步微调基模型,再测试模型的检测效果。实验检测结果如表 1 所示。为了避免实验的偶然性误差,每个评估实验进行了 10 次,最后取其平均值。

所提方法在训练过程有 2 个优化目标:1)交叉熵





(a) 本方法微调之前



(b) 本方法微调之后

图 3　本方法微调模型前后的能量分数分布变化

表 1　模型微调前后检测异常样本的对比实验

正常样本	异常样本	FPR 95		AUROC		AUPR	
		基模型	+所提方法	基模型	+所提方法	基模型	+所提方法
CIFAR-10	Texture	52.46 ± 0.88	26.52 ± 0.84	85.54 ± 0.30	94.07 ± 0.32	95.63 ± 0.13	98.44 ± 0.12
	SHVN	27.72 ± 0.57	17.57 ± 0.74	94.64 ± 0.11	96.96 ± 0.19	98.80 ± 0.03	99.36 ± 0.05
	Places365	43.13 ± 0.80	27.38 ± 0.99	87.73 ± 0.37	93.26 ± 0.31	96.36 ± 0.16	98.19 ± 0.11
	LSUN-C	09.36 ± 0.64	4.38 ± 0.52	98.05 ± 0.07	99.09 ± 0.08	99.58 ± 0.02	98.81 ± 0.02
	LSUN Resize	25.33 ± 0.79	9.23 ± 0.76	94.96 ± 0.16	98.12 ± 0.12	98.82 ± 0.05	99.61 ± 0.03
	iSUN	29.93 ± 0.86	12.74 ± 0.66	93.86 ± 0.20	97.56 ± 0.10	98.57 ± 0.06	99.48 ± 0.03
	Mean	31.32	16.30	92.46	96.51	97.96	99.15
CIFAR-100	Texture	80.44 ± 0.61	68.33 ± 0.68	74.91 ± 0.27	81.85 ± 0.29	93.18 ± 0.11	95.38 ± 0.12
	SHVN	74.75 ± 0.67	56.05 ± 0.52	84.42 ± 0.24	89.32 ± 0.28	96.55 ± 0.07	97.68 ± 0.09
	Places365	82.23 ± 1.03	73.11 ± 0.89	74.70 ± 0.47	79.71 ± 0.49	93.25 ± 0.17	94.75 ± 0.21
	LSUN-C	43.01 ± 1.37	33.02 ± 0.83	92.60 ± 0.21	94.11 ± 0.16	98.42 ± 0.05	98.75 ± 0.04
	LSUN Resize	71.15 ± 1.30	49.61 ± 0.66	81.42 ± 0.37	88.36 ± 0.25	95.13 ± 0.14	97.14 ± 0.10
	iSUN	75.92 ± 0.89	54.93 ± 0.72	79.55 ± 0.52	87.40 ± 0.18	94.66 ± 0.21	96.92 ± 0.07
	Mean	71.25	55.84	81.26	86.79	95.20	96.77

损失保证模型对正常样本的分类能力不会受到影响；2)稀疏优化损失指导模型增加正常样本特征向量的稀疏程度。稀疏的特征具有更低的能量值，拉开了正常样本与异常样本之间的能量值差异，从而提高了能量分数检测算法的性能。表 1 为正常样本模型在微调前后，检测不同的异常样本的表现情况。由表 1 可知，经过所提方法微调后的模型，检测异常样本的能力得到全面提升。总之，在 CIFAR-10 模型上，所提方法将 6 个数据集上 FPR95 平均降低了 15.02%；在 CIFAR-100 模型上，FPR95 平均降低了 15.41%；这是一个非常显著的进步，所提方法无需辅助数据集。

3.4.2　对比不同的方法

上述实验对比了模型在不同数据集上的具体表现，说明了所提方法的有效性，但是仅看有效性还不能突出所提方法在当前的研究领域中的优势。实验继续对比当前比较流行的检测方法，包括 MSP[5]、ODIN[7]、Mahalanobis[9] 及 EBD[6]。评估所有方法的模型都是基于 WideResNet 网络训练的分类模型。

为了体现实验对比的公平性，所有方法都未用到辅助数据集，只需正常样本训练数据集。表 2 为不同方法的检测表现情况。表 1 中的数据均为模型在 6 个测试数据集上进行了 10 次实验后结果的平均值。可以发现，所提方法在所有检测指标上均领先于其他方法，很大程度提高了当前方法的检测性能，说明所提方法非常具有竞争力。

表 2　不同方法检测异常样本的表现情况

分类模型	方法	FPR 95	AUC	AUPR
CIFAR-10 WideResNet	MSP	52.36	90.42	97.42
	ODIN	36.40	91.86	98.03
	Mahalanobis	38.08	91.37	98.52
	EBD	31.32	92.46	97.96
	本方法	16.30	96.51	99.15
CIFAR-100 WideResNet	MSP	80.55	75.43	93.32
	ODIN	76.65	77.34	94.14
	Mahalanobis	60.05	80.36	95.07
	EBD	71.25	81.26	95.20
	本方法	55.84	86.79	96.77

4　结束语

提出了一种简单且有效的损失函数，用于对已经训练好的 CNN 分类模型进行微调，显著提高了模型检测异常样本的能力。其优化的目标是增加正常样本特征的稀疏性，从而使得正常样本的能量分数更低，进一步增加正常样本和异常样本之间的可区分程度。实验结果表明，包括正常样本采用不同的 CNN 网络，模型经过简单微调后，一致提升了检测不同类型异常样本的能力。所提方法实现了即使不暴露异常数据给模型，也能达到与之相当的效果。下一步将继续实验所提方法在更多不同 CNN 网络上的表现情况及研究方法的可迁移性，探索在更高维度和更大规模数据集上的表现。

参考文献：

[1] HE Kaiming，ZHANG Xiangyu，REN Shaoqing，et al. Delving deep into rectifiers：surpassing human-level performance on imagenet classification［C］//Proceedings of the IEEE International Conference on Computer Vision. Piscataway，NJ：IEEE Press，2015：1026-1034.

[2] NGUYEN A，YOSINSKI J，CLUNE J. Deep neural networks are easily fooled：high confidence predictions for unrecognizable images［C］//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE Press，2015：427-436.

[3] HEIN M，ANDRIUSHCHENKO M，BITTERWOLF J. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem［C］//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE Press，2019：41-50.

[4] FILOS A，TIGKAS P. Can autonomous vehicles identify，recover from，and adapt to distribution shifts?［C］//Proceedings of the 37th International Conference on Machine Learning. New York，NY：ACM Press，2020，119：3145-3153.

[5] HENDRYCKS D，GIMPEL K. A baseline for detecting misclassified and out-of- distribution examples in neural networks［EB/OL］.（2016-10-07）［2021-10-03］. https://arxiv.org/abs/1610.02136.

[6] LIU Weitang，WANG Xiaoyun，OWENS J，et al. Energy-based out-of-distribution detection［C］//Advances in Neural Information Processing Systems 33. Cambridge，MA：MIT Press，2020：21464-21475.

[7] LIANG Shiyu，LI Yixuan，SRIKANT R. Enhancing the reliability of out-of-distribution image detection in neural networks［EB/OL］.（2017-07-08）［2020-08-30］. https://arxiv.org/abs/1706.02690.

[8] HSU Y C，SHEN Yilin，JIN Hongxia，et al. Generalized ODIN：detecting out-of-distribution image without learning from out-of-distribution data［C］//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE Press，2020：10951-10960.

[9] LEE K，LEE H，SHIN J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks［C］//NeurIPS 2018. Cambridge，MA：MIT Press，2018：7167-7177.

[10] DAN H，MANTAS M，THOMAS D. Deep anomaly detection with outlier exposure［EB/OL］.（2018-12-04）［2022-01-28］. https://arxiv.org/abs/1812.04606.

[11] TORRALBA A，FERGUS R，FREEMAN W T. 80 million tiny images：a large data set for nonparametric object and scene recognition［J］. IEEE Transactions on Pattern Analysis and Machine Intelligence，2008，30（11）：1958-1970.

[12] PAPADOPOULOS A，RAJATI M R，SHAIKH N，et al. Outlier exposure with confidence control for out-of-distribution detection［J］. Neurocomputing，2021，441：138-150.

[13] YU Q，AIZAWA K. Unsupervised out-of-distribution detection by maximum classifier discrepancy［C］//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway，NJ：IEEE Press，2019：9518-9526.

[14] LEE K，LEE H，LEE K，et al. Training confidence-calibrated classifiers for detecting out-of-distribution samples［EB/OL］.（2017-11-26）［2022-02-23］. https://arxiv.org/abs/1711.09325.

[15] CHEN J F，LI Y X，WU X，et al. Informative outlier matters：robustifying out-of-distribution detection using outlier mining［C］//ECML PKDD 2021：Machine

Learning and Knowledge Discovery in Databases. Berlin,German:Springer,2021,129:8-26.
[16] LI Y, VASCONCELOS N. Background data resampling for outlier-aware classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020:13218-13227.
[17] MOHSENI S,PITALE M,YADAWA J,et al. Self-supervised learning for generalizable out-of-distribution detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2020,34:5216-5223.
[18] YANG Jingkang,WANG Haoqi,FENG Litong,et al. Semantically coherent out-of-distribution detection [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway,NJ:IEEE Press,2021:8301-8309.
[19] LECUN Y, CHOPRA S, HADSELL R, et al. Predicting Structured Data[M]. MA:MIT Press,2006.
[20] KRIZHEVSKY A,HINTON G. Learning multiple layers of features from tiny images[D]. Toronto: University of Toronto,2009:48-60.
[21] ZHOU B,LAPEDRIZA A,KHOSLA A,et al. Places:a 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,40(6):1452-1464.
[22] NETZER Y,WANG T,Coates A,et al. Reading digits in natural images with unsupervised feature learning [C]//NIPS Workshop on Deep Learning and Unsupervised Feature Learning. MA:MIT Press. 2011.
[23] MIRCEA C,SUBHRANSU M,IASONAS K,et al. Describing textures in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 3606-3613.
[24] GIRSHICK R,DONAHUE J,DARRELL T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE Press,2014:580-587.
[25] YU F, SEFF A, ZHANG Y. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop[EB/OL]. (2015-06-10)[2021-06-04]. https://arxiv. org/abs/1506. 03365.
[26] ZAGORUYKO S, KOMODAKIS N. Wide residual networks[J]. arXiv preprint arXiv:1605. 2016:07146.

实习编辑:高波